ORIGINAL ARTICLE

The Power of Noise and the Art of Prediction

ZhiMin Xiao[1]* | **Steve Higgins**[2]

[1, 2]School of Education, Durham University

Correspondence
School of Education, Durham University, Durham, DH1 1TA, UK
Email: zhimin.xiao@durham.ac.uk

Funding information
This research was funded by a grant to Durham University from the Education Endowment Foundation in England.

Data analysis usually aims to identify a particular signal, such as an intervention effect. Conventional analyses often assume a specific data generation process, which implies a theoretical model that best fits the data. Machine learning techniques do not make such an assumption. In fact, they encourage multiple models to compete on the same data. Applying logistic regression and machine learning algorithms to real and simulated datasets with different features of noise and signal, we demonstrate that no single model dominates others under all circumstances. By showing when different models shine or struggle, we argue that it is important to conduct predictive analyses using cross-validation for better evidence that informs decision making.

KEYWORDS
Cross-Validation; Evidence-Based Policy; *K*-NN; Logistic Regression; Prediction; Random Forests

1 | TWO MODELLING APPROACHES

Data analysis is usually about identifying signal from noise. But data, particularly social science data, can be truly noisy, partly because the outcome is often a human construct, which can only be measured with some error. Noise can also stem from other factors, such as the collection of data on variables that are not correlated with the outcome of interest, or the addition of interaction and/or higher order terms, which can easily fail in-sample goodness-of-fit tests

(Breiman, 2001), meaning the combination of many variables and their various transformations in a conventional, say, linear regression, can be so flexible or fit the observed data so well that it has little value in the explanation of new or out-of-sample data. The noise mentioned above can be minimised in careful research designs and sound data analyses. Nevertheless, theories about best designs and views on best analysis strategies can be another source of noise, because the best approach to data analysis for a given study often differs in theory from person to person, even for those who are from the same discipline (Xiao et al., 2016). One classical example is a statistical phenomenon called Lord's Paradox (Lord, 1967, 1969), where the relationship between two variables change in both magnitude and direction when a third variable is statistically controlled for (Holland, 2005; Tu et al., 2008a,b; Wainer, 1991; Wainer and Brown, 2004; Xiao et al., 2017a). In education, decisions on not just variables but also types of models are often necessary, in order to account for the fact that pupils are usually nested within classes, which are from schools located in different regions. Depending on the structure of a specific dataset, such choices can result in considerable differences in point estimates and uncertainties surrounding those estimates (Xiao et al., 2016, 2017a,b).

Moreover, single best models, or the practices of using just one model and explaining why that model is the best according to a mathematical theory or evidence found elsewhere, cannot be statistically compared unless some of them are nested within others (Shmueli and Koppius, 2011). Donoho called the analytical approach that relies on a single best model derived from a mathematical formula "generative modeling" (2015), where a data generation process is assumed and a single best model, which must exist because of the assumptions made, is then deployed to analyse the data. But the choice of that model can itself be a source of variation in results because of the theoretical differences mentioned above. Consequently, the model may lead to "irrelevant theory and questionable scientific conclusions" (Breiman, 2001) because it is usually more about a data generation and selection process than about how the real world functions or the underlying problem to be solved. When published, the results may further justify the choice of the theoretically best model in subsequent studies, particularly when they are linked with research funding streams, which in turn make the results more salient or more noticeable in the literature. This feedback loop can be pernicious (O'Neil, 2016), if policy decisions made on the evidence from a single best model produce unintentional and undesirable consequences (Merton, 1936) to those who participated in the studies and/or beyond.

Most regression models we see in educational research are of the generative type, which can be theoretically best

because of the asymptotic guarantee, meaning if an intervention were to be repeated many times until all samples in a population are exhausted, the model is guaranteed to predict the correct outcome. This sounds reassuring, but in reality, we do not live in an "asymptopia" (Domingos, 2012), an imaginary situation where an intervention can be replicated many times without any constraint. This implies that, if model A is better than model B given infinite data, due to bias-variance trade-off or balance between precision and uncertainty, there is no guarantee that the former will be better than the latter given finite data or a particular dataset. As such, we also need "predictive modeling" (Donoho, 2015; Hofman et al., 2017), which is generally agnostic about a data generating mechanism and focuses on how well it predicts the future rather than how well it fits the data after the fact (Popkin, 2015). Typically, predictive modelling encourages multiple models to learn from and work on multiple datasets, some of which are used to train the models, others put aside as test sets, just as we turn a ball many times and each time we make a prediction about the patterns on the side we do not see using the information on the side we can see. The performances of the trained models are then judged against a common task, usually, predictive accuracy on test sets, which is easy-to-understand and can be compared across datasets and over time (Breiman, 2001; Donoho, 2015; Hofman et al., 2017; James et al., 2015).

In education, when predictions are made on learning outcomes, it has the spirit of predictive modelling, which relies on some observed data in the past (training sets) and its performance is assessed on how accurately it predicts yet-to-be-observed outcomes (test sets). Note that predictive models do not have to be sophisticated. Assuming we know nothing or little about the past, guessing is one predictive model, which is usually less accurate than the averaging of past outcomes as another predictive model. As we gather more data, we can employ more powerful models such as regression to make more accurate predictions. Predictive modelling thus embraces changes in the real world and always improves as we feed it with more data (Popkin, 2015).

Predictive modelling also provides timely feedback for analysts to assess how successful the tools they have deployed actually are in the wild. As a result, the best performing models can be efficiently utilised for real-world applications, which can then enhance the roles evidence has to play in decision making and reduce the gap between research and practice (Shmueli, 2010). This approach overcomes one problem of single best models, where different analysts analyse the same dataset in their own manner and may produce different results and make different claims about the performance of their preferred methods (Breiman, 2001; Donoho, 2015; Hand, 2006; Xiao et al., 2016). If

we do not know which, if any, of the best models actually worked because of the problems associated with in-sample strength-of-fit measures, such as the coefficient of determination or R^2 in a linear regression (Breiman, 2001; Shmueli and Koppius, 2011), the conclusions drawn from the results of single best models may be too dependent on error or noise, making them effectively just noises themselves. This only adds to the challenge of evidence-informed policy and practice by confusing decision makers with varying advice.

In many social science studies, such as the educational interventions funded by the Education Endowment Foundation (EEF) in the UK, predictive modelling is yet to be widely appreciated (Shmueli and Koppius, 2011), despite the aforementioned advantages and its rapid development and application in other fields (DeRubeis et al., 2014; Kapelner et al., 2014; Kennedy et al., 2017; Popkin, 2015; Hofman et al., 2017; Tetlock et al., 2017), such as the engineering of personal computing and smartphones in our daily lives, and individualised or precision medicine in scientific research. For the time being, the design and analysis of EEF trials focus primarily on average treatment effect on the treated, which is helpful if we are only interested in the mean. But one technical difficulty of the approach is that many large-scale EEF trials are producing negligibly small effect sizes, which has many implications, including an ethical one of randomly assigning students to a group that we could have predicted to be non-optimal for specific subgroups of students, given the evidence gathered earlier about a trial and the data collected about the students under concern. Also, when an intervention does not specifically target a subgroup, such as Free School Meal (FSM) pupils in England, an estimate of effect is always reported for the group with caution, which is analogous to saying: this intervention has such an effect on FSM pupils, but the public should not really trust the result. Subgroup analysis is notoriously difficult (Assmann et al., 2000; Lagakos, 2006; Petticrew et al., 2012; Song and Bachmann, 2016; Wang et al., 2007), but it is a step towards personalised intervention effect, which can be computed using predictive modelling and has the potential to solve the challenges associated with average treatment effect and subgroup analysis.

The process of randomly splitting data into training and test sets can transform the technical procedure of generative modelling into that of its predictive counterpart. In other words, we can also use data to train conventional regression models and then employ the trained models to predict outcomes on the test set. However, the assumptions made by conventional linear regression and new machine learning techniques we are going to encounter shortly are very different. The former assumes a linear and mathematical relationship between the outcome and observed features

of the studied, whereas the latter takes data as the only input and allows the data to tell what that relationship really is (Popkin, 2015). Making connections between the "old" and "new" thus avoids defying the inferential contributions generative modelling has made.

To find out when some models shine and others struggle, we apply conventional logistic regression and some machine learning techniques to real and simulated datasets. We use logistic regression because the outcome is binary and it is easier to understand the percentage of accurate predictions than the mean squared errors when the outcome is continuous in an ordinary least squares regression. Nevertheless, the logic is the same, be it a classification or regression problem. The machine learning techniques compared are K-Nearest Neighbours (K-NN) and random forest, which are known to perform well in prediction without making untenable assumptions often associated with conventional regression methods. In this study, K-NN refers to a method that first tries to identify a certain number of observations in the training set that are most similar to an observation it aims to predict the outcome for in the test set. Since the outcome is binary, it adopts a majority vote in the training set as the outcome for the observation in the test set. Clearly, the predictive accuracy of the method depends on the number of neighbours chosen in the training set, hence the name K-NN (Hastie et al., 2013; James et al., 2015). Random forest uses a random subset of covariates in a dataset to split the data until the terminal nodes or leaves of individual decision trees consist of very similar observations. Using a random subset rather than all of the covariates to conduct binary cuts of the data helps grow diverse trees based on bootstrapped re-sampling of the observed data. In the long run, the prediction is less biased and more reliable, as the method prevents the same strong predictors from re-occurring in all the subsets. As in K-NN for classification, prediction is made on an observation in the test set using the most commonly occurring outcome of training observations in the same region (Hastie et al., 2013; James et al., 2015).

2 | THE FALLACY OF MORE DATA

In this study, we first show that measurement error in some outcome of interest can mask the relationship between covariates and the outcome. In the simulation, we suppose the outcome is an unweighted mean of two covariates,

which, together with the outcome and other variables, come from a normal distribution with the same mean of zero and standard deviation one. To add measurement error into the outcome, we introduce a normal distribution with the same mean but different standard deviations, which represent the strength of the noise in the outcome. Figure 1 visualises the effect of measurement error on the correlations between the outcome and the two signal variables, X_1 and X_2. As the noise gets louder and louder, the correlation coefficients between the outcome and the two signal variables become smaller and smaller. This change is not affected by the change in sample size, which suggests that the common focus on larger sample sizes to detect an intervention effect might be inappropriate when the outcome measure is prone to errors. For educational interventions, such as those funded by the EEF, it is therefore crucial to choose the right test as an outcome measure. Otherwise, the real effect of an intervention might never be detected, even when the sample size in a given study is large, if there is substantial measurement error in the test (see also Loken and Gelman, 2017). An example of this can be found in the Building Blocks intervention in the US, which might have suffered from some measurement error, where the large-scale pre-K program focused on skills such as counting that children will eventually master as they grow, even without the intervention (Mervis, 2017). To avoid this risk, the EEF normally requires independent evaluators to pre-specify a standardised national test, which minimises the effect of the noise in measurement at either baseline or post-test.

Noise does not simply exist as measurement error. It can also exert its power through covariates, particularly when there are interactions among them. In social science research, there is usually a tendency to collect data on as many covariates as possible. This is partly due to competing theories that inform data collection. As an example, the association under the miasma theory in the mid-1800s between cholera and personal habits and characteristics, such as strong emotions like fear and immorality, specifically overindulgence in alcohol and sex, predispositions often linked at the time with the "lower classes" (Tulodziecki, 2011). But it also has something to do with the more the merrier philosophy, according to which, more information on as many variables as possible, at worst, provides no extra information about the outcome. However, one risk of the practice may be that the benefits of having more data on irrelevant covariates can be sometimes outweighed by "the curse of dimensionality" (Domingos, 2012), meaning the noise in the data is so strong that it swamps the signal we are looking for, or at least, it makes that search very inefficient.

Again, we use simulations to illustrate the power of noise, particularly when there are interaction terms. Suppose

the outcome this time is a categorical variable with two possible values of one and zero in an educational intervention. When it takes on the value of one, it means a student gained from the intervention. Otherwise, the student made no progress from baseline to post-test. At the beginning of the intervention, we also had a number of baseline measures, among which there might exist interactions. First, let's suppose an interaction exists between two covariates. We implement this interaction effect by randomly shifting half of the observations in the covariates up by a certain amount, and the other half down by the same amount. We also adjust the outcomes accordingly so that observations that were randomly shifted up are more likely to take on the value of one, and those shifted down less likely to be one. Apart from the two covariates that are interacted, there are also others that are neither interacted with one another nor correlated with the outcome variable. Those null variables are in effect noises, which vary in number from dataset to dataset, as the earlier measurement errors vary in strength from one simulation to another. To illustrate the effect of interaction in the simulated data, we have produced some pairwise scatter plots. As shown in Figure 2, when the interaction effect, which is determined by the amount randomly shifted up or down earlier, is tiny, as in (a), the observations in any pair of variables are randomly scattered. But, as we increase the interaction effect by increasing the shifted amount, the outcomes become increasingly separable and the interaction effect is clearer.

To analyse the simulated datasets, we introduce a number of analytical models, which are logistic regression, random forests, and K-NN with K taking on varying number of values (James et al., 2015). To show how stable those analytical models are, we also simulate each specification three times, which results in three performance outcomes for each of the models mentioned above. As shown in Figure 3, the performance of logistic regression is no better than tossing a coin when there is just one interaction term, which involves two interacted variables. Random forests have higher predictive power when the sample sizes are larger. K-NN outperform others when the values of K are appropriate in a given simulation. When sample size is 100, the most appropriate K is between 5 and 10, as sample size increases to 500 or 1000, the most appropriate K is about 100. However, the results are much more stable when the sample size is 1000 across the three simulated datasets.

The scenario described above involves only one interaction term (or two interacted covariates) in the simulated dataset. Now let's see how the models perform when we increase the power of noise and signal by adding more null and interacted variables. As we can see in Figure 3, when there are two interacted variables, random forests can almost

match K-NN at the most appropriate value of K. Logistic regression again crumbles, regardless of the strength of signal and noise in the simulated data. Nevertheless, when the number of interacted variables climbs up to five, the only model that can achieve acceptable predictive accuracy (or low test error) is K-NN at the right level of K, of which the choice is much narrower than when there are only two interacted variables.

In the above-simulated datasets, the conventional logistic regression is no better than guesswork even when a single interaction term that involves two interacted covariates exists. This does not suggest that it is no use at all. For the logistic regression to have higher predictive power, we can add into the regression an interaction term, which will substantially improve the predictive accuracy of the model. The addition of the interaction term is straightforward when we know which two variables are interacted and when the total number of covariates is not large so that we can explore all possible pairwise combinations of covariates in the data. But when there are many variables, as is the case in many social science studies, it will be practically impossible to exhaust all possible combinations. Usually, analysts of social research data add interaction terms when a theory suggests them what covariates are likely to interact with one another. If serious interaction exists but analysts fail to address it accordingly, the logistic regression will produce evidence that is inadequate to inform decision making. The so-called evidence would be just another source of noise in the literature. Given that we normally do not know how many interaction terms exist in a dataset and it is prohibitive to examine all the combinations when there are as few as ten covariates, it is no surprise that machine learning techniques such as random forests and K-NN are increasingly perceived to be better models for data analysis. As it becomes easier and perhaps less expensive to collect more data, these techniques are likely to appeal to more and more social scientists in the years to come. But they also have limitations, as the above simulations show, when there are five interacted covariates or more, random forests are no better than logistic regression.

3 | THE FALLACY OF FREE LUNCHES

The simulations described above are, after all, only simulations. The performances of these models may change when they are tested on a new dataset with different features. In this section, we use a dataset that is openly available to the

public and well-known to the machine learning community (LeCun et al., 1998). The dataset has 60,000 observations in the training set and 10,000 in the test set. Each observation represents a hand-written digit ranging from 0 to 9, and there are 784 columns, each of which contains values that represent degrees of grey. To see which model has the highest level of prediction accuracy in reading the digit 8 in the test set, we use observations from the training set to train candidate models, which are then used to predict the outcome in the test set. In the training, we also alter the number of sample sizes in each re-sample, so that we can observe how the models learn from the data. The models trained are logistic regression, decision trees, bagging, and random forests (James et al., 2015; Liaw and Wiener, 2002). We exclude K-NN for this dataset because it is known to have high predictive accuracy in reading hand-written digits (Domingos, 2012) and the main purpose of using this dataset is to show the effect of input knowledge on the performances of conventional and machine learning techniques. The last three of the learners selected are related, because each decision tree is a hierarchy of cuts using either continuous or categorical variables, and bagging refers to bootstrapped aggregation, where each bootstrapped re-sample is used to grow a decision tree. Since bagging uses all features in each re-sample, and variables that are highly correlated with the outcome are likely to be selected first, the algorithm will produce very similar trees in the end. As introduced earlier, random forests overcome this problem by randomly choosing a subset of features in each bootstrapped re-sample of the data. As a result, the trees grown in the model will be very diverse, and the average performance of the forests is usually better than that of bagging (James et al., 2015). However, the algorithm of random forests runs much faster than that of bagging. A useful metaphor to explain the difference in speed is that it is like randomly opening a subset of drawers in a chest of drawers that contain different pieces of information about the outcome, rather than all the drawers each time.

Figure 4 (b) shows the performances of the four models mentioned above. As we can see, when the training sample size is 300, the accuracy level of logistic regression is slightly above 50%. However, when the sample size increases to about 1000, logistic regression can achieve about 80% accuracy, which slowly increases as more samples are used to train the model. Nevertheless, it is no match for the other three machine learning techniques, particularly when the training size is small. This dataset is unique in the sense that there are 784 covariates, when the training size is below 1000, logistic regression really struggles in pulling the accuracy level up to those of its counterparts, which have at least 90% accuracy even when sample size is as low as 300. Their performances, particularly those of random forests,

increase as more and more training data are fed into them. However, it is worth mentioning that the computational costs for bagging and logistic regression are remarkably high, although they can eventually achieve comparable accuracy levels of random forests. The performance of decision trees also increases as training samples increase, but it is less accurate than bagging and random forests.

As we have demonstrated, developing models with high predictive accuracy requires a lot of "black art" (Domingos, 2012), which can rarely be found in statistical textbooks. Although conventional logistic regression can eventually achieve a similar level of accuracy to that of random forests in the above case, the latter can get more from less and it runs much faster than the former. The results reported above also show that machine learning techniques cannot do magic without input knowledge. There is no such thing as a free lunch. But they can, as with a lever, turn a small amount of input into a large amount of output, but this obviously varies from model to model (Domingos, 2012).

So far, we have used simulated and real datasets to test and compare the performances of both logistic regression and machine learning techniques such as decision trees, bagging, and random forests. When there are more variables than observations in a dataset and there are very few interaction terms, the performance of random forests is truly impressive in terms of predictive accuracy. Unlike logistic regression and bagging, it is not computationally expensive to run. Moreover, it does not require analysts to fine tune many parameters, such as the values of K in K-NN. Unsurprisingly, it appeals to more and more analysts.

4 | THE FALLACY OF SINGLE BEST MODELS

Next, we use a few more datasets from large-scale educational interventions funded by the EEF in England. Unlike the datasets we have seen so far, EEF datasets are highly curated and structured. In three of the four cases that follow, participants were randomly assigned to intervention or control groups, and the tests used in the trials were standardised national tests in England, suggesting that the measurement error is likely to be low. As in earlier datasets, observations are randomly split into training and test sets, the models are first trained in the training set, and then tested for their predictive accuracy, test error, sensitivity, and specificity. Four out-of-sample performance metrics are used this time

because the costs associated with different misclassification errors can be different (Hand, 2006). Since there are many ways to split the data into training and test sets, we report the results from two splitting methods, the first rows use 68% of the observed samples in the interventions to train the models and the rest as test sets. The second rows result from bootstrapped training and test sets, meaning sample sizes in the training sets are equal to the observed sample sizes of the interventions. However, some observations will be selected more than once for the training sets while others will not be selected at all, which are then used as test sets. As expected, the two rows in Figure 5 (a–d) have almost identical results across all the metrics. This is ideal because we do not want the results to be sensitive to different cross-validation methods. The four datasets also vary in sample and effect sizes. Using zero gain as the cut-off, the outcome takes on either one or zero, meaning either progress or no progress in the tests at the end of the interventions.

Before we look at the results, we will provide some background information about the trials. The metacognition intervention, called ReflectED, aimed to improve pupils' ability to think about and manage their own learning (Motteram et al., 2016). It is a school-based randomised controlled trial with randomisation at class level. In the final analysis, the study involved 1507 pupils from 30 schools, and the primary outcome was age standardised mathematics score. Chess in Schools is an intervention that randomly allocated 100 schools (4009 pupils) to either intervention or control. Intervention schools taught children how to play chess over a year, whereas control schools were business-as-usual. The primary outcome was Key Stage 2 mathematics score one year after the intervention (Jerrim et al., 2016). Improving Writing Quality is a smaller intervention with a large effect size. It involved 261 pupils from 23 primary schools, which were randomly allocated to receive training on writing or to continue with business-as-usual (Torgerson et al., 2014). Unlike the first three, Tutor Trust Secondary is a quasi-experimental design, which matched 781 participating pupils with 100,991 others in a comparison group who did not participate in the small group or one-to-one tutoring intervention but had similar demographic and socio-economic characteristics (**?**). The outcome chosen for this study is performance on GCSE mathematics.

Figure 5 shows the performances of three models, which are logistic regression, random forests, and K-NN with varying number of neighbours in the training set used for prediction on the test set. As for predictive accuracy in (a), random forests have the most impressive performance, which is followed by logistic regression. K-NN, across all values of K, do not perform as well as expected. However, when K is 100, it has the highest level of specificity, meaning when

the outcome is zero, the model accurately predicts zero with the highest level of accuracy. This model is therefore less likely to produce false positives than others for this intervention. In terms of sensitivity, logistic regression and random forests are about equal. When K is 100, K-NN has the lowest level of sensitivity, which means it has the lowest level of predictive accuracy when the outcome is one and it accurately predicts one. This model is thus more likely to produce false negatives. In sub-figure (b), logistic regression and random forests perform about equally well in terms of accuracy and test error. However, the former has the highest level of specificity and the latter the highest level of sensitivity. K-NN, again, is far behind the first two across all metrics of performance. In (c), the performance of logistic regression is better than any other model considered across all the metrics. Random forests closely follow that of logistic regression. Given the large sample size and imbalanced structure of the data in (d), random forests and K-NN truly shine in all aspects.

As we can see, across the four trial datasets, higher levels of accuracy always correlate with lower levels of test error, but higher levels of sensitivity do not always imply lower levels of specificity. The patterns thus suggest that it is important to compare model performances across multiple metrics. Besides, random forests, while impressive when sample size is relatively large, are not necessarily better than conventional logistic regression. When the sample size is relatively small and the data is clean and well structured, the widely perceived superior machine learning technique cannot outperform its conventional counterpart. However, this does not imply that more experienced users of random forests cannot fine tune its parameters to "squeeze" the best performance out of it (Domingos, 2012; Hand, 2006).

5 | CONCLUSION

Taken together, we have demonstrated that no single model dominates others under all circumstances. Most studies using single best models explain why their models are the best, but say little or nothing about how or why their predictions or inferences might be wrong (Subrahmanian and Kumar, 2017). To demonstrate the risks of taking this approach, we have shown both. For instance, logistic regression is no better than tossing a coin when sample sizes are small, and when there are more covariates than observations, not to mention the roles interaction terms may play.

Random forests are no better than logistic regression when there are many interaction terms and when the noise swamps the signal. *K*-NN at the appropriate value of *K* can achieve the lowest level of test error, even when there are many interaction terms and null variables. However, they are far behind logistic regression and random forests when the data are "overly clean" (Shmueli, 2010) and the observations are very similar across intervention and control groups, in which case, *K*-NN at all levels of *K* simply make random predictions. The findings from all the datasets used in the study thus present a compelling case that single best models cannot be known a *a priori* and it is crucial to cross-validate results and compare model performances using multiple metrics.

This is a troubling time for evidence-informed policy, partly because we do not always agree on what constitutes as the best evidence, but it also stems from the fact that the path from knowledge to power is not always linear. Evidence is just one ingredient that goes into the policy mix (Malakoff, 2017). In order to present the best possible evidence, the conclusions made above are ever more important. As we do not live in an asymptopia and sometimes decisions have to be made in a timely fashion, we can no longer safely say at the end of an intervention that the findings are mixed, therefore more studies are needed. When an answer straddles both sides of "maybe", it precludes accountability (Tetlock et al., 2017). So, we suggest that one way forward is to concede our exclusive reliance on generative modelling, which risks producing research results that may have little relevance to practice. Although the theoretically best approach can have high in-sample explanatory power or breadth, it does not necessarily follow that its out-of-sample predictive power will be precise (Trafimow and Uhalt, 2015). Therefore, it is important to make sure that research findings from social science research such as the educational interventions in this study can explain the causal mechanism well, but also have sufficient predictive quality (Shmueli, 2010), or give us some idea, in advance, of what impact an intervention will have, for whom and where (Clauset et al., 2017). Otherwise, the gap between methodological advance and practical application will be widened and the path of evidence to impact becomes even more winding.

CONFLICT OF INTEREST

We have no conflicts of interest to disclose.

REFERENCES

Assmann, S. F., Pocock, S. J., Enos, L. E. and Kasten, L. E. (2000) Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet*, **355**, 1064–1069.

Breiman, L. (2001) Statistical Modeling: The Two Cultures. *Statistical Science*, **16**, 199–231.

Clauset, A., Larremore, D. B. and Sinatra, R. (2017) Data-driven predictions in the science of science. *Science*, **350**, 477–480.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A. and Lorenzo-Luaces, L. (2014) The personalized advantage index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS ONE*, **9**, 1–8.

Domingos, P. (2012) Tapping into the "folk knowledge" needed to advance machine learning applications. *Communications of the ACM*, **55**, 78–87.

Donoho, D. (2015) 50 Years of Data Science. In *Tukey Centennial Workshop*, 1–41. Princeton. URL: `http://goo.gl/9gWEQp`.

Hand, D. J. (2006) Classifier Technology and the Illusion of Progress. *Statistical Science*, **21**, 1–14.

Hastie, T., Tibshirani, R. and Friedman, J. (2013) *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer, 2 edn.

Hofman, J. M., Sharma, A. and Watts, D. J. (2017) Prediction and explanation in social systems. *Science*, **355**, 486–488.

Holland, P. W. (2005) Lord's Paradox. In *Encyclopedia of Statistics in Behavioral Science* (eds. B. S. Everitt and D. C. Howell), vol. 2, 1106–1108. John Wiley & Sons.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2015) *An Introduction to Statistical Learning: with Applications in R*, vol. 64. New York: Springer, 6th edn.

Jerrim, J., Macmillan, L., Micklewright, J., Sawtell, M. and Wiggins, M. (2016) Chess in Schools. URL: `http://goo.gl/HSf1Gv`.

Kapelner, A., Bleich, J., Levine, A., Cohen, Z. D., DeRubeis, R. J. and Berk, R. (2014) Inference for the Effectiveness of Personalized Medicine with Software. *arXiv*. URL: `http://arxiv.org/abs/1404.7844`.

Kennedy, R., Wojcik, S. and Lazer, D. (2017) Improving election prediction internationally. *Science*, **355**, 515–520.

Lagakos, S. W. (2006) The challenge of subgroup analyses - Reporting without distorting. *New England Journal of Medicine*, **354**, 1667–1669.

LeCun, Y., Cortes, C. and Burges, C. J. (1998) The MNIST database of handwritten digits. URL: http://goo.gl/yF0yH.

Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News*, **2**, 18–22.

Loken, E. and Gelman, A. (2017) Measurement error and the replication crisis. *Science*, **355**, 584–585.

Lord, F. M. (1967) A paradox in the interpretation of group comparisons. *Psychological Bulletin*, **68**, 304–305.

— (1969) Statistical adjustments when comparing preexisting Groups. *Psychological Bulletin*, **72**, 336–337.

Malakoff, D. (2017) A Matter of Fact. *Science*, **355**, 562–563.

Merton, R. K. (1936) The Unanticipated Consequences of Purposive Social Action. *American Sociological Review*, **1**, 894–904.

Mervis, J. (2017) No easy answers: What does it mean to ask whether a prekindergarten math program "works"? *Science*, **355**, 568–571.

Motteram, G., Choudry, S., Kalambouka, A., Hutcheson, G. and Barton, A. (2016) ReflectED. URL: http://goo.gl/JX9oGX.

O'Neil, C. (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group.

Petticrew, M., Tugwell, P., Kristjansson, E., Oliver, S., Ueffing, E. and Welch, V. (2012) Damned if you do, damned if you don't: subgroup analysis and equity. *Journal of Epidemiology & Community Health*, **66**, 95–98.

Popkin, G. (2015) A twisted path to equation-free prediction. *Quanta Magazine*. URL: https://goo.gl/QWGt3J.

Shmueli, G. (2010) To explain or to predict? *Statistical Science*, **25**, 289–310.

Shmueli, G. and Koppius, O. R. (2011) Predictive Analytics in Information Systems Research. *MIS Quarterly*, **35**, 553–572.

Song, F. and Bachmann, M. (2016) Cumulative subgroup analysis to reduce waste. *BMC Medicine*, **14**, 1–8.

Subrahmanian, V. S. and Kumar, S. (2017) Predicting human behavior: The next frontiers. *Science*, **355**, 489–489.

Tetlock, P. E., Mellers, B. A. and Scoblic, J. P. (2017) Bringing probability judgments into policy debates via forecasting tournaments. *Science*, **355**, 481–483.

Torgerson, D., Torgerson, C., Ainsworth, H., Buckley, H., Heaps, C., Hewitt, C. and Mitchell, N. (2014) Improving Writing Quality. URL: http://goo.gl/YBuVsg.

Trafimow, D. and Uhalt, J. (2015) The alleged tradeoff between explanatory breadth and predictive power. *Theory & Psychology*, **25**, 833–840.

Tu, Y.-K., Baelum, V. and Gilthorpe, M. S. (2008a) A structural equation modelling approach to the analysis of change. *European Journal of Oral Sciences*, **116**, 291–296.

Tu, Y.-K., Gunnell, D. and Gilthorpe, M. S. (2008b) Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon–the reversal paradox. *Emerging themes in epidemiology*, **5**, 1–9.

Tulodziecki, D. (2011) A case study in explanatory power: John Snow's conclusions about the pathology and transmission of cholera. *Studies in History and Philosophy of Biological and Biomedical Sciences*, **42**, 306–316.

Wainer, H. (1991) Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, **109**, 147–151.

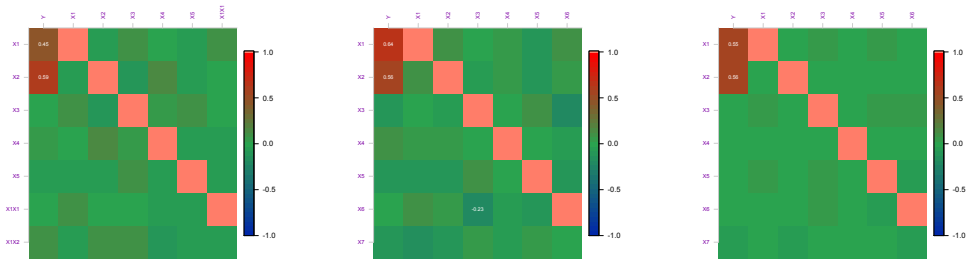
Wainer, H. and Brown, L. M. (2004) Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *The American Statistician*, **58**, 117–123.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J. and Drazen, J. M. (2007) Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. *New England Journal of Medicine*, **357**, 2189–2194.

Xiao, Z., Higgins, S. and Kasim, A. (2017a) An Empirical Unravelling of Lord's Paradox. *The Journal of Experimental Education*. URL: `https://doi.org/10.1080/00220973.2017.1380591`.

— (2017b) Seeing is Believing: Impact Visualisation in Educational Interventions. *SocArXiv*. URL: `https://osf.io/preprints/socarxiv/e4prv/`.

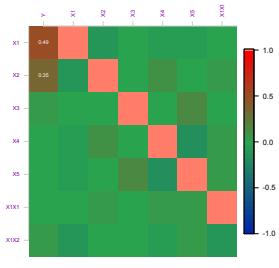
Xiao, Z., Kasim, A. and Higgins, S. (2016) Same difference? Understanding variation in the estimation of effect sizes from educational trials. *International Journal of Educational Research*, **77**, 1–14.



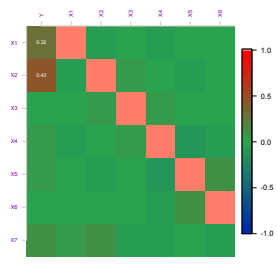
(a) N=200, NV=5, SD=0.5

(b) N=200, NV=8, SD=0.5

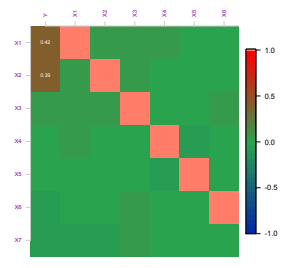
(c) N=1000, NV=8, SD=0.5



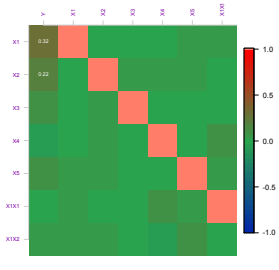
(d) N=200, NV=5, SD=1



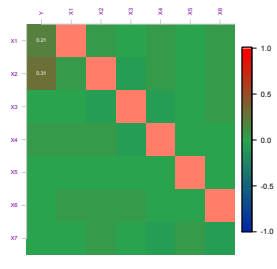
(e) N=500, NV=8, SD=1



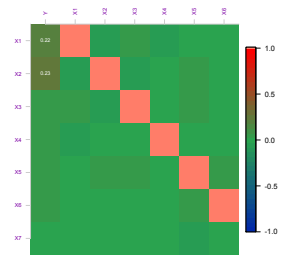
(f) N=1000, NV=18, SD=1



(g) N=500, NV=5, SD=2

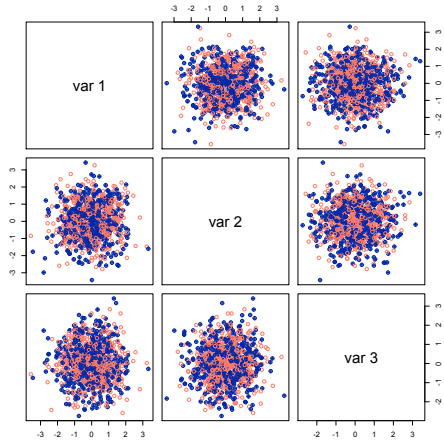


(h) N=800, NV=10, SD=2

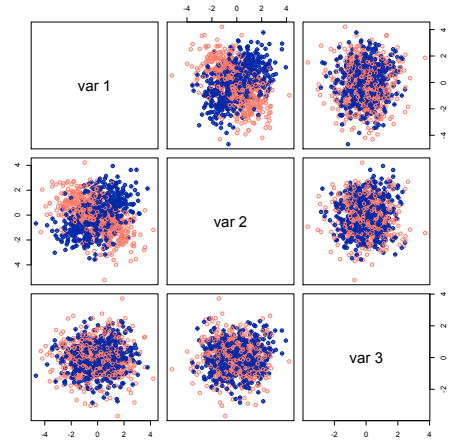


(i) N=1000, NV=10, SD=2

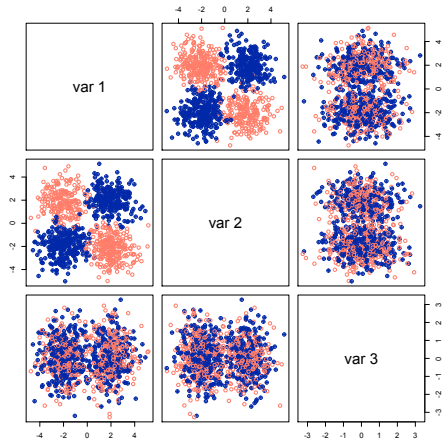
FIGURE 1 **Simulated impact of measurement error.** Correlation matrices between outcome Y, signal variables X_1, X_2, and other null variables, which are weakly correlated with the outcome. The greener the cells, the weaker the correlations. In the first row, the standard deviation of the noise in outcome is 0.5, which increases to 1 in the middle row and 2 in the bottom row. Note the first two cells of the Y column become greener and greener as the noise gets louder and louder. N represents simulated sample size. NV is the number of variables, and SD is the standard deviation of the noise. For a better visualisation effect, each matrix plots the correlations between the first eight variables of the corresponding simulation only.



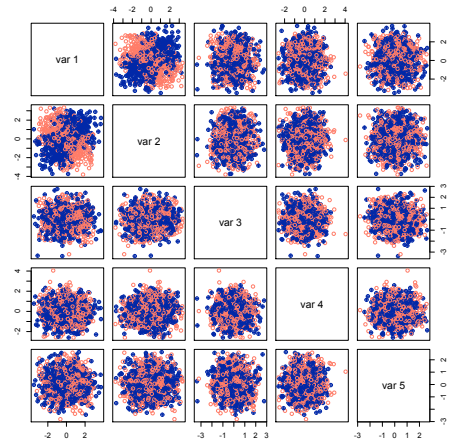
(a) N = 1000, Vs = 3, D = 0.1



(b) N = 1000, Vs = 3, D = 1

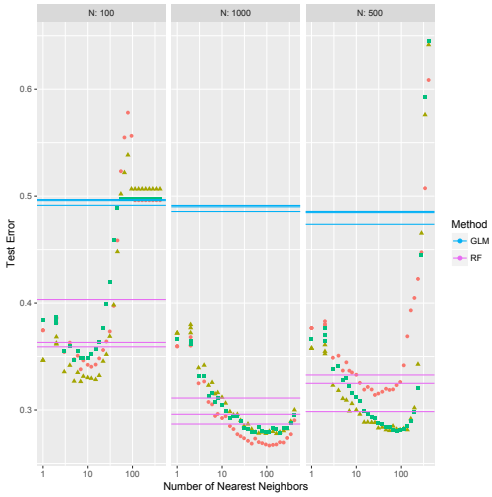


(c) N = 1000, Vs = 3, D = 2

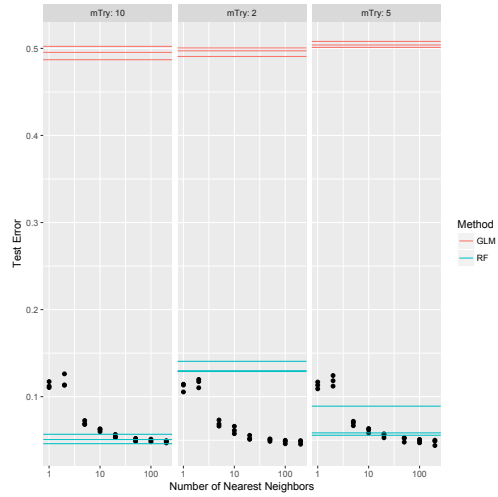


(d) N = 1000, Vs = 5, D = 1

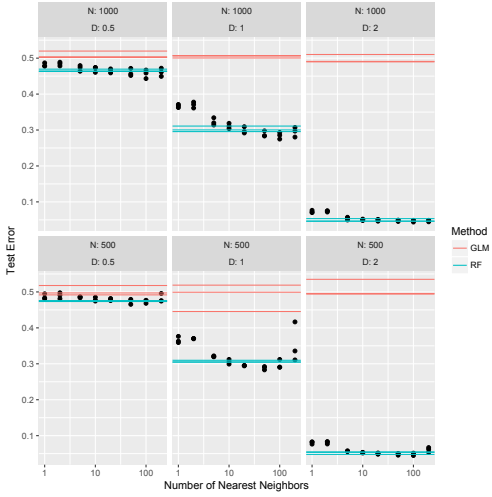
FIGURE 2 **Visualising interaction effect.** The four simulated datasets have the same sample size (N) of 1000, and in each case, only the first two variables are interacted. However, the datasets differ in the number of variables (Vs) and degree of separability in outcome (D), namely, interaction effect.



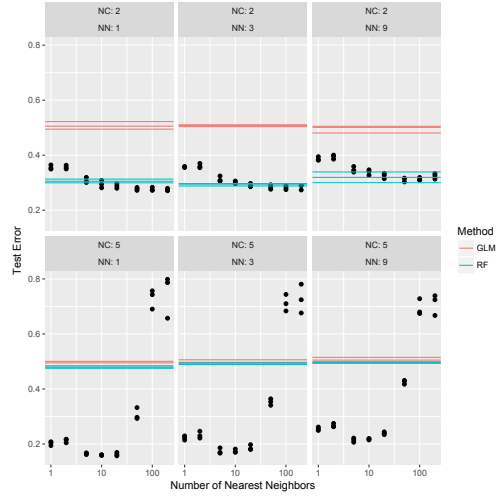
(a) N=(100, 500, 1000), NC=2, NN=3, D=1



(b) N=2500, NC=2, NN=18, D=2

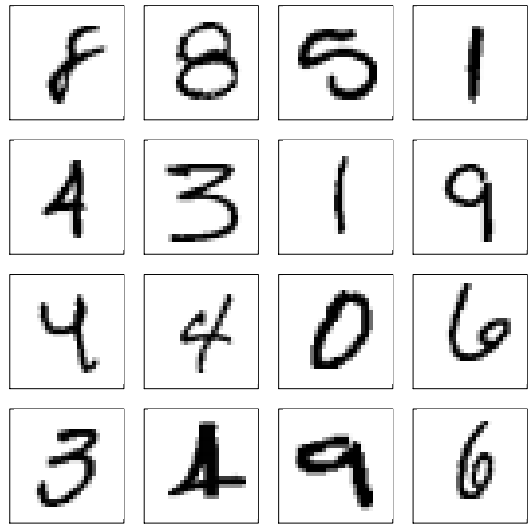


(c) N=(500, 1000), NC=2, NN=3, D=(0.5, 1, 2)

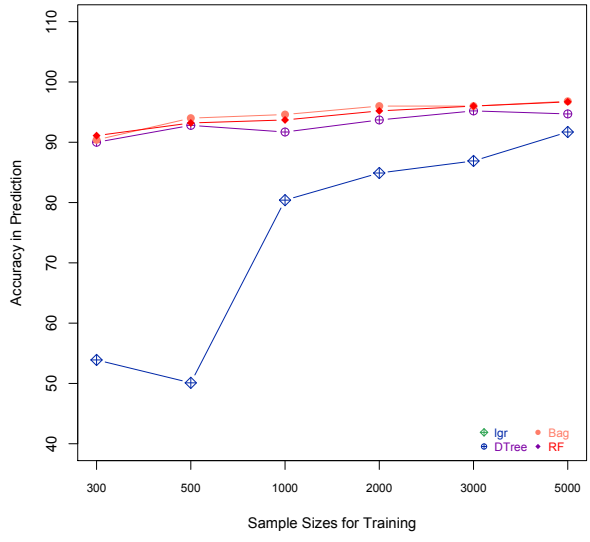


(d) N=1000, NC=(2, 5), NN=(1, 3, 9), D=(1, 2)

FIGURE 3 **Analysing simulated datasets with interactions.** Performances of different models on simulated datasets with different sample sizes (N) and interaction effects (D), varying numbers of null (NN) and interacted variables (NC). Sub-figure (b) shows the effect of change in the number of random predictors (`mTry`) used in each re-sample for random forests. The simulations use only one performance metric, which is prediction error on the test set.

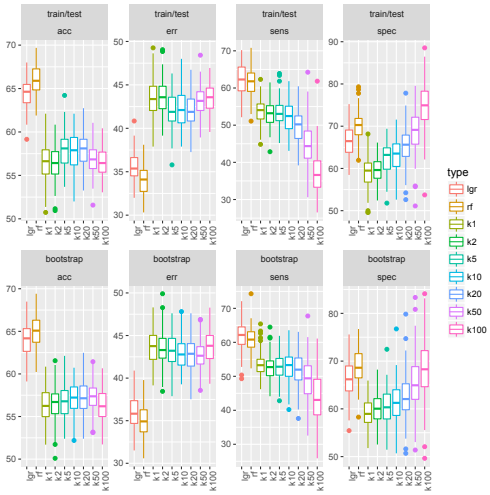


(a) 16 random digits

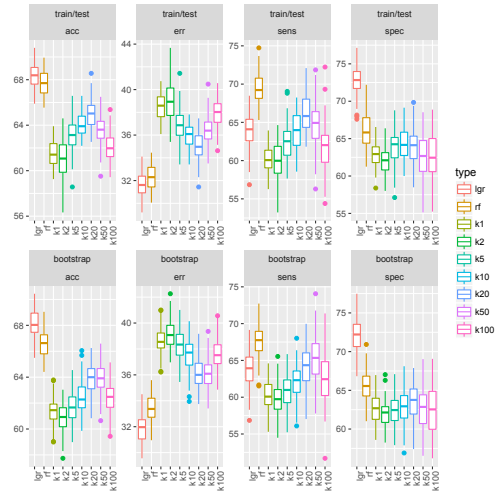


(b) Effect of input knowledge

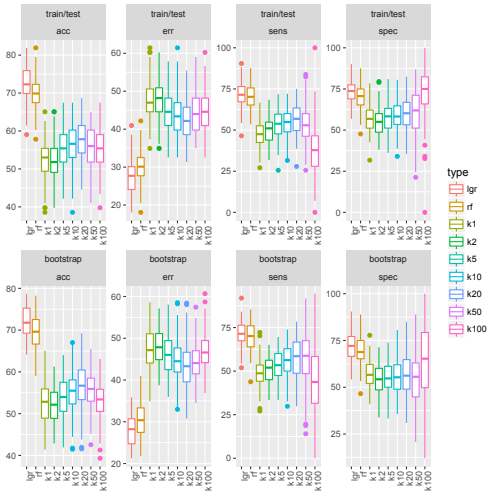
FIGURE 4 **Training models to read hand-written digits.** Sub-figure (a) is a random reading of 16 hand-written digits from the training set. Each digit is located in the center of a 28 by 28 grid, which forms a row with 784 columns if the cells in the grid are stacked up to form just one row. When there are 10,000 rows in the test set, there are 10,000 digits. Sub-figure (b) shows the effect of sample sizes in the training set on the prediction accuracy of the four models, logistic regression (`lgr`), decision trees (`DTree`), bagging (`Bag`), and random forests (`RF`).



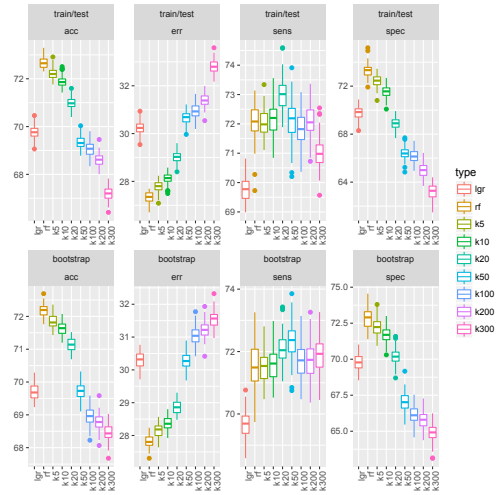
(a) ReflectED: −0.01(−0.35, 0.32)



(b) Chess in Schools: −0.05(−0.23, 0.14)



(c) Improving Writing Quality: 0.67(0.07, 1.28)



(d) Tutor Trust Secondary: 0.08(−0.22, 0.39)

FIGURE 5 Performances of multiple models on EEF data. There are four performance metrics, prediction accuracy (`acc`), prediction error (`err`), sensitivity (`sens`), and specificity (`spec`). `train/test` and `bootstrap` represent two different data splitting methods. `lgr` and `rf` refer to logistic regression and random forests, respectively. All models with a letter `k` in the label are K-NN with different values of K. Project titles are followed by effect estimates and their 95% confidence intervals.